AI Fundamentals of Counterfactuals

Computational Foundations of Counterfactuals

Plan for the Day

CF Tutorial

TIME	Topics
9:00 AM	***Introduction***
	Hello and Introducing Ourselves!
	Hands-on: *Trying Our Study (follow link)*
9:30 AM	***Historical Fundamentals of Counterfactuals***
	From Philosophy to XAI (via Psychology)
	Two Sample User Studies and Q&A
10:30 AM	COFFEE (10:30-11:00)
11:00 AM	***Fundamentals of Counterfactuals in AI***
	Formalisation
	Modelling Approaches & Key Constraints
11:30 AM	***Using Counterfactual Algorithms***
	Hands-on: *A Counterfactual Toolbox (AA)*
	Hands-on: *Checking Out Notebooks and Q&A*
12:00 PM	***Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design
12:30 PM	LUNCH (12:30-14:00)
2:00 PM	***Algorithmic Growth Points***
	Computational Future Directions and Q&A
2:30 PM	***More Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design (cont.)
3:00 PM	COFFEE (15:00-15:30)
3:30 PM	***From Fundamentals to an Actual User Study***
	User Studies II: A More Complex Design
	User Studies III: Even More Complex Designs
	Hands-on: *Looking At Our Study*
5:00 PM	***Closing Session, Discussion and Final Q&A***
	TUTORIAL END

You Are Here!

Detailed Outline

STEP 1 Formalization

STEP 2 Modeling Approaches

STEP 3 Key Constraints

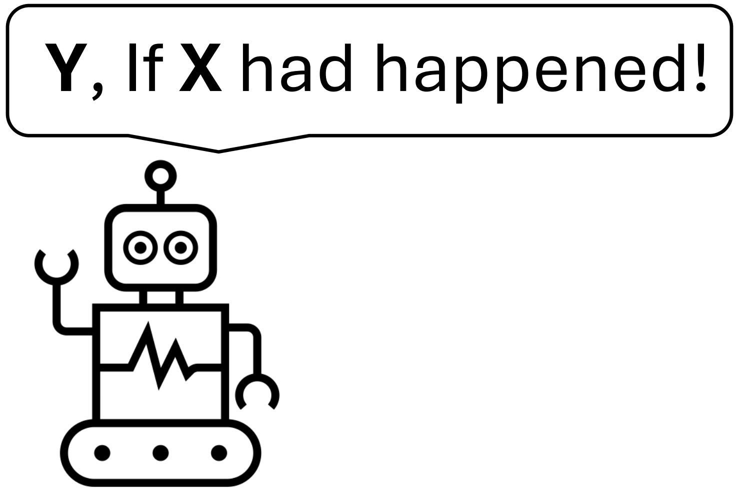
STEP 4 Hands-On: Counterfactuals in Python

Formalization of Counterfactuals

- Two properties:

1. Contrastive
2. Proximity/Closeness/Cost
   - Often a p-norm – but is this realistic?

     => Domain specific!



Modeling Approaches

- Optimization problem [Wachter 2017]:

$$\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min} \boxed{\ell\left(h(\vec{x}_{\text{cf}}), y_{\text{cf}}\right)} + C \cdot \boxed{\theta(\vec{x}_{\text{cf}}, \vec{x}_{\text{orig}})}$$

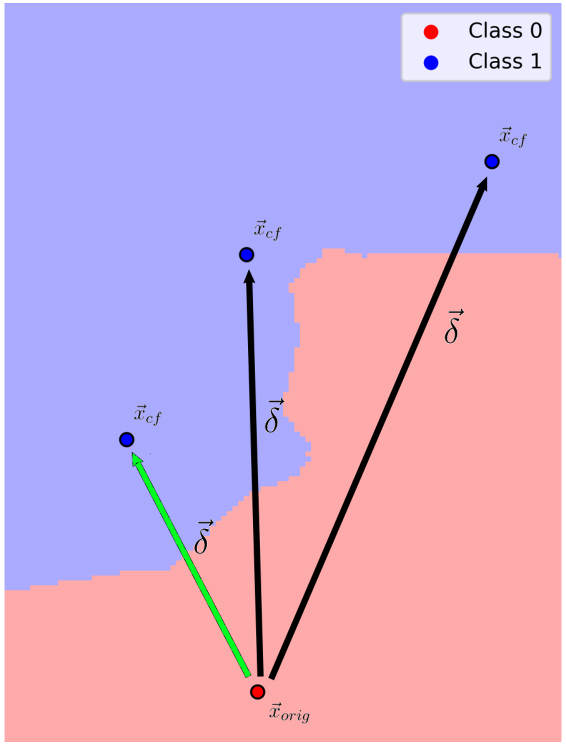
Contrastive **Cost/Proximity**

$$\text{Change } \vec{\delta} = \vec{x}_{\text{cf}} - \vec{x}_{\text{orig}}$$

In constraint form:

$$\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min} \boxed{\theta(\vec{x}_{\text{cf}}, \vec{x}_{\text{orig}})} \quad \text{s.t.} \quad \boxed{h(\vec{x}_{\text{cf}}) = y_{\text{cf}}}$$

Cost/Proximity **Contrastive**

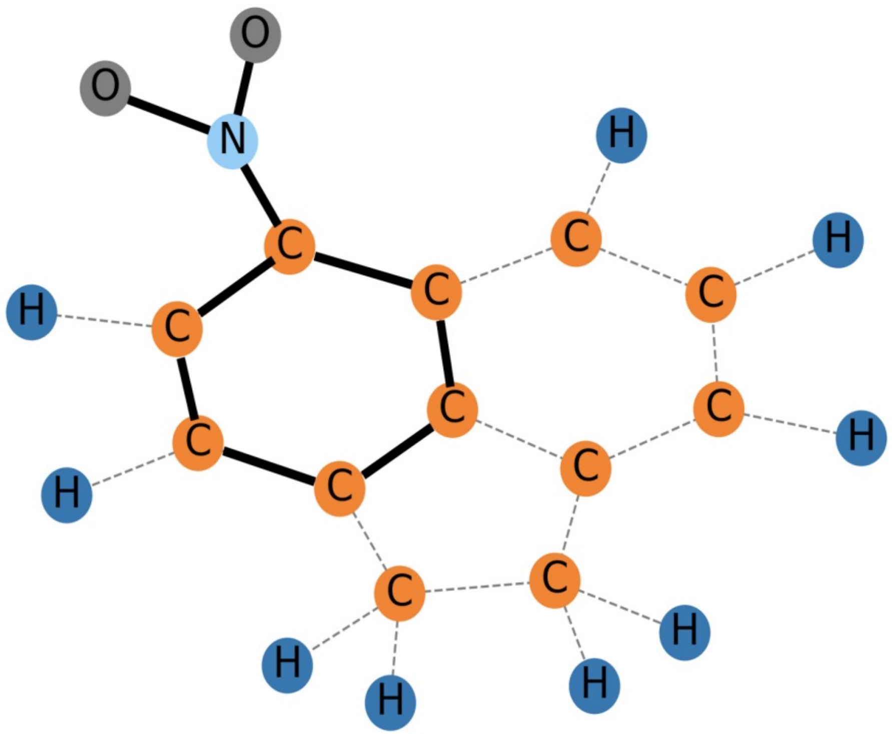


Modeling Approaches

- How to solve the optimization problems?

$$\underset{\vec{x}_{\text{cf}} \in \mathbb{R}^d}{\arg\min}\ \ell\big(h(\vec{x}_{\text{cf}}), y_{\text{cf}}\big) + C \cdot \theta(\vec{x}_{\text{cf}}, \vec{x}_{\text{orig}})$$

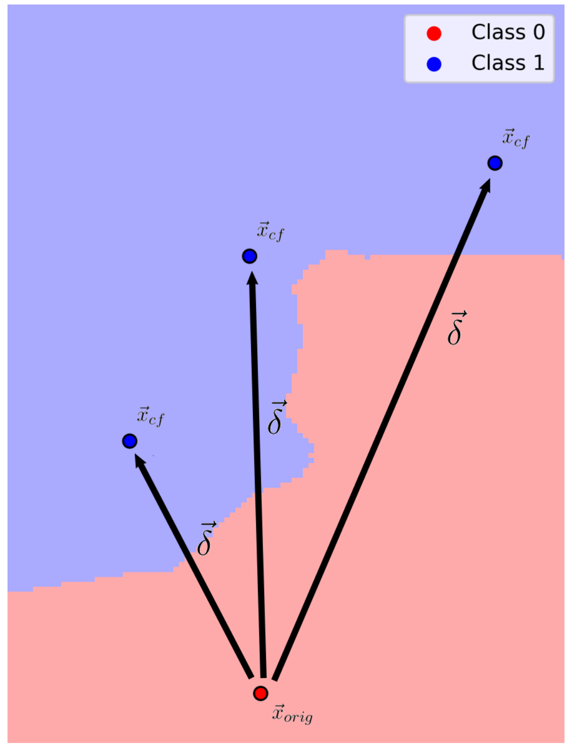
  - Black-box solvers
  - Gradient descent
  - Domain-specific methods (images, text, graphs, etc.)

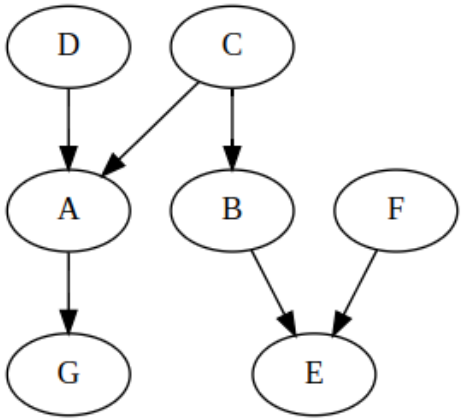


Things to keep in Mind

- Uniqueness?
  => "Rashomon Effect"
- What about Causality? [Karimi 2021]
  => Often, feature independence is assumed!

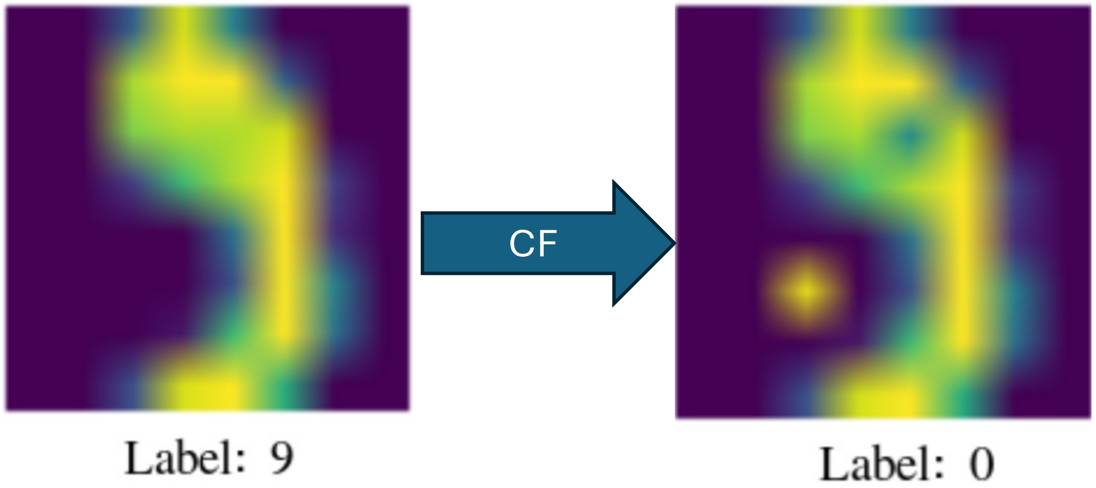
Incorporate Causal models!





Fundamentals & Key Constraints

- Contrastivity and Cost/Proximity are not enough!
  - e.g. CF != Adversarial
- Other important aspects:
  - Plausibility
  - Actionability
  - Diversity
  - ...

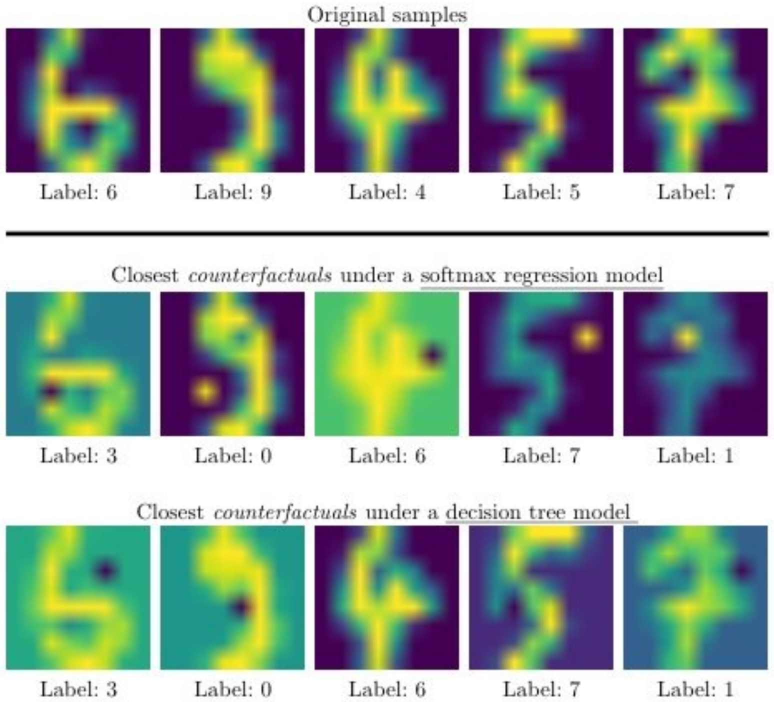


Plausibility & Actionability

- Problem: Classic CFs [Wachter 2017] often adversarials

  => Not useful in practice! [Smyth 2022]

- **CFs must be plausible and actionable**



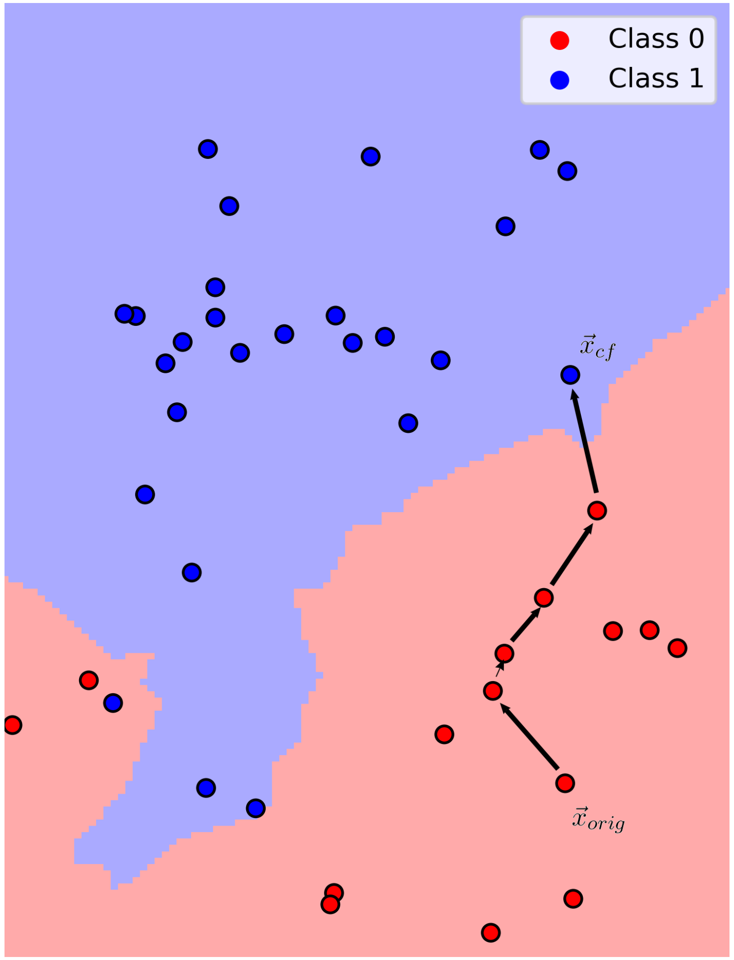
Plausibility & Actionability

- Two popular & general approaches:
  - FACE [Poyiadzi 2020]
  - Density constraints [Artelt 2020]

$$\underset{\vec{x}_{\mathrm{cf}} \in \mathbb{R}^d}{\arg\min} \; \theta(\vec{x}_{\mathrm{cf}}, \vec{x}_{\mathrm{orig}})$$

$$\text{s.t.} \;\; h(\vec{x}_{\mathrm{cf}}) = y_{\mathrm{cf}}$$

$$\hat{p}_{y_{\mathrm{cf}}}(\vec{x}_{\mathrm{cf}}) \geq \underbrace{\delta}_{\text{lower-bound on density}}$$



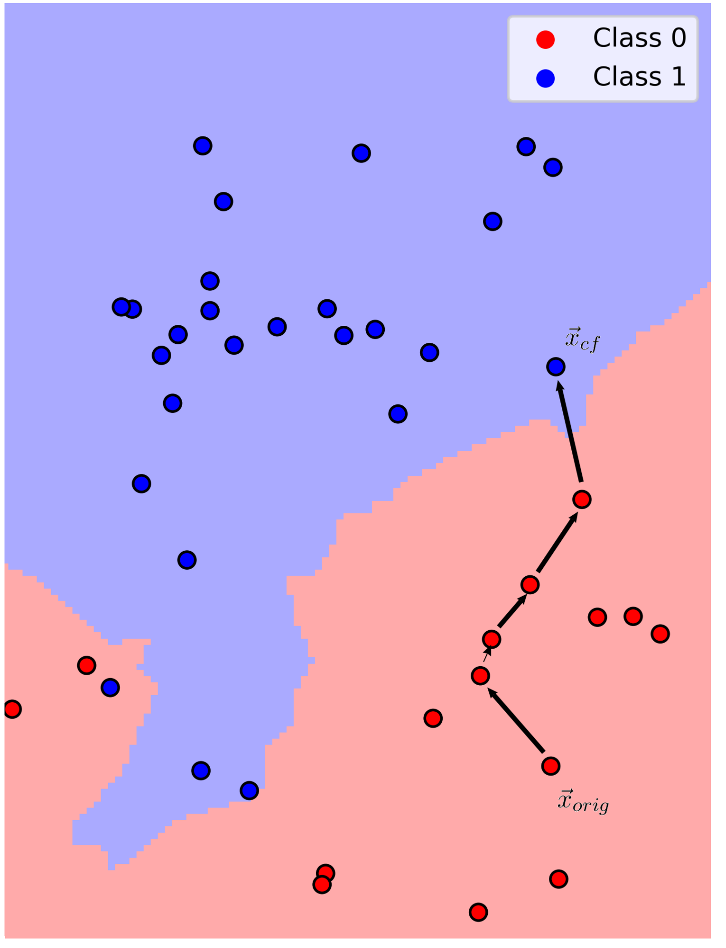
FACE: Feasible, Actionable CFs

[Poyiadzi 2020]

- Sequence of changes
- Limit feasible set to observed data
  - "No feature independence"
  - Actionable & Plausible

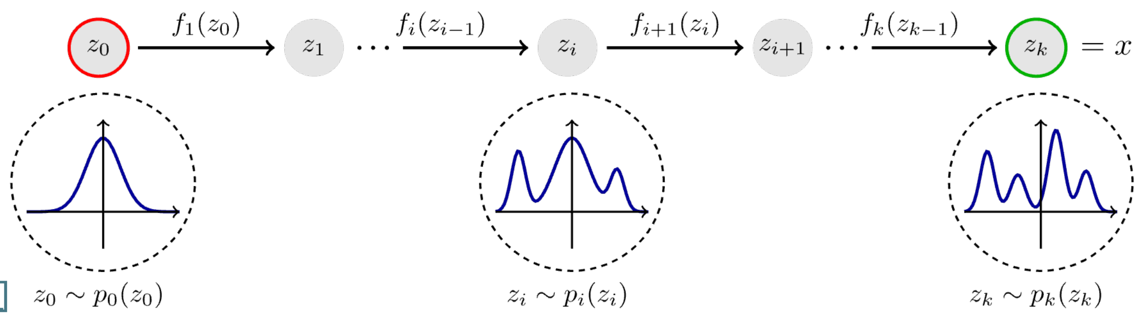
- <u>Graph:</u> KDE, k-NN, or ϵ-graph

=> Find shortest path



Density Constraints

- **CF in high-density region**
- Methods:
  - GMMs [Artelt 2020]
  - KDE [Förster 2021]
  - GANs [Mertes 2022]
  - Normalizing Flows [Wielopolski 2024]



$$\underset{\vec{x}_{\mathrm{cf}} \in \mathbb{R}^d}{\arg\min}\ \theta(\vec{x}_{\mathrm{cf}}, \vec{x}_{\mathrm{orig}})$$

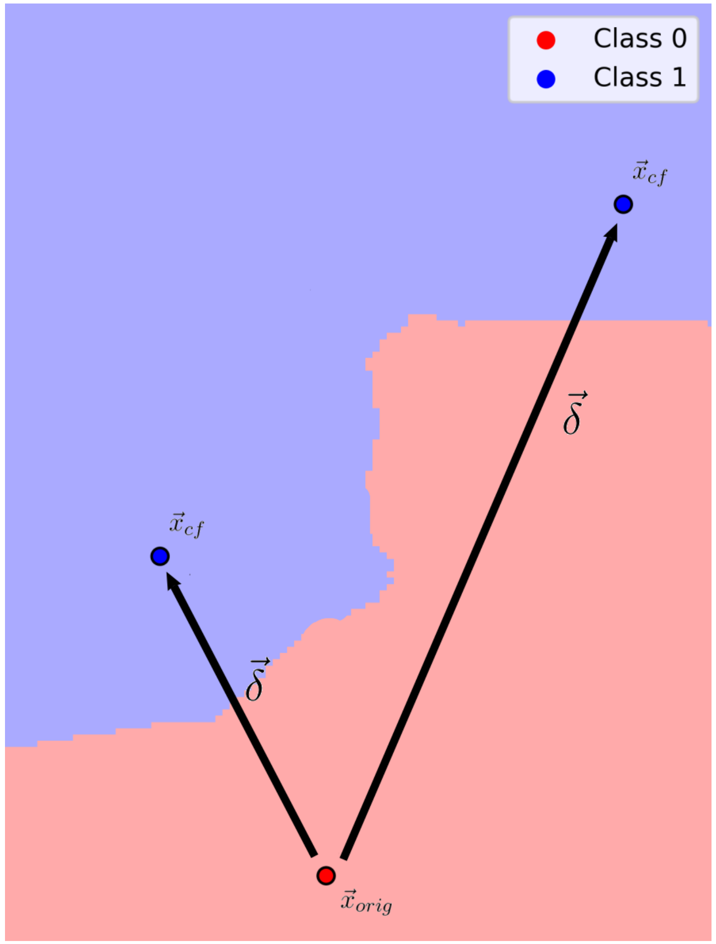
$$\text{s.t.}\quad h(\vec{x}_{\mathrm{cf}}) = y_{\mathrm{cf}}$$

$$\hat{p}_{y_{\mathrm{cf}}}(\vec{x}_{\mathrm{cf}}) \geq \underbrace{\delta}_{\text{lower-bound on density}}$$

Diversity

- Missing uniqueness
  => "Rashomon Effect"
- Generate multiple "different" CFs
  - Let the user choose!
  - More robust? [Leofante 2024]

RQ: How to model diversity?



Diversity

- Yet another constraint:
  - Quantify difference between CFs (e.g. cost, common features, etc.)

- Toolbox: DiCE [Mothilal 2020]

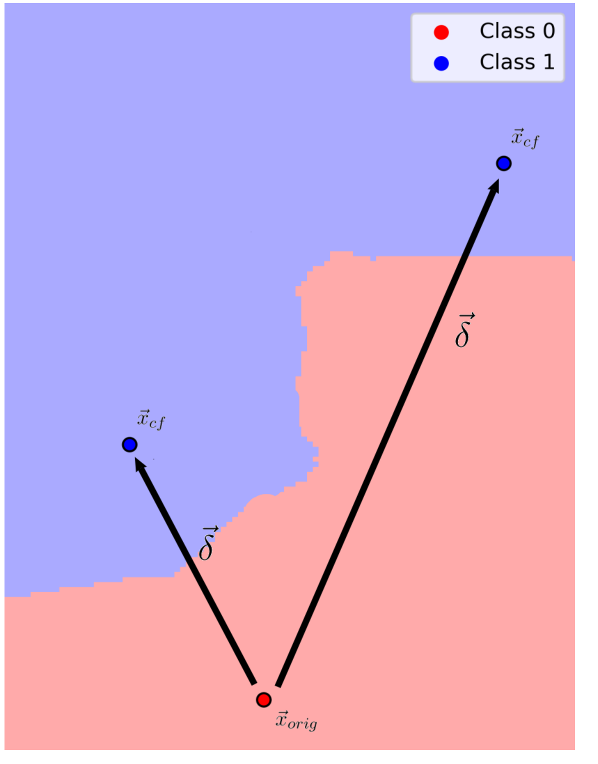
=> See Hands-on session

$$\underset{\{\vec{x}_{\mathrm{cf}_i} \in \mathbb{R}^d\}}{\arg\min} \underbrace{\sum_i \ell\big(h(\vec{x}_{\mathrm{cf}_i}), y_{\mathrm{cf}}\big)}_{\text{Constrasting}} + \underbrace{C_1 \sum_i \theta(\vec{x}_{\mathrm{cf}_i}, \vec{x}_{\mathrm{orig}})}_{\text{Complexity}}$$

$$\boxed{\underbrace{+\, C_2 \sum_{i,j} \psi\left(\vec{x}_{\mathrm{cf}_i}, \vec{x}_{\mathrm{cf}_j}\right)}_{\text{Diversity}}}$$

Summary & Conclusion

- Two important properties:
  - Contrastive
  - Cost/Proximity
- Key constraints:
  - Plausibility & Actionability
  - Diversity



Hands-On: Counterfactuals in Python

https://tinyurl.com/6mc35pya

Plan for the Day

CF Tutorial

TIME	Topics
9:00 AM	***Introduction***
	Hello and Introducing Ourselves!
	Hands-on: *Trying Our Study (follow link)*
9:30 AM	***Historical Fundamentals of Counterfactuals***
	From Philosophy to XAI (via Psychology)
	Two Sample User Studies and Q&A
10:30 AM	COFFEE (10:30-11:00)
11:00 AM	***Fundamentals of Counterfactuals in AI***
	Formalisation
	Modelling Approaches & Key Constraints
11:30 AM	***Using Counterfactual Algorithms***
	Hands-on: *A Counterfactual Toolbox (AA)*
	Hands-on: *Checking Out Notebooks and Q&A*
12:00 PM	***Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design
12:30 PM	LUNCH (12:30-14:00)
2:00 PM	***Algorithmic Growth Points***
	Computational Future Directions and Q&A
2:30 PM	***More Fundamentals of User Studies***
	User Studies I: A Simple Two-Group Design (cont.)
3:00 PM	COFFEE (15:00-15:30)
3:30 PM	***From Fundamentals to an Actual User Study***
	User Studies II: A More Complex Design
	User Studies III: Even More Complex Designs
	Hands-on: *Looking At Our Study*
5:00 PM	***Closing Session, Discussion and Final Q&A***
	TUTORIAL END

You Are Here!

Other & Current Topics

- **Semi-Factuals**
- Fairness
- Robustness
- Manipulations
- Group Counterfactuals

Semi-Factual Explanations

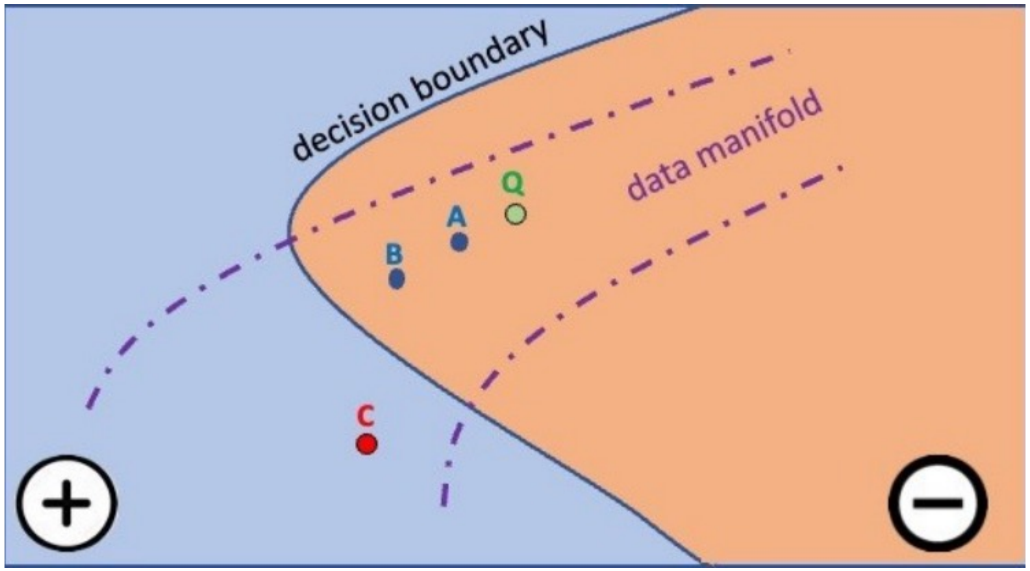
- "Even-If" Explanations [Aryal 2023]

- "**Even if** you used twice as much fertilizer last month, the crop yield ***would still have been*** **the same**"

- Why Semi-Factuals?
  - Potentially have major impact (decreasing causal link)
  - May be useful for positive outcomes
  - Have "good" emotional impacts

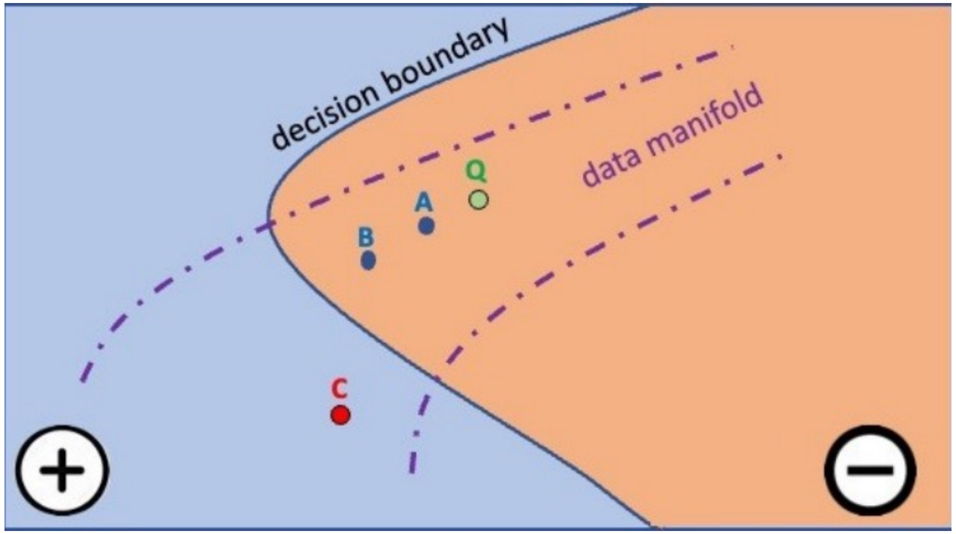
Semi-Factual Explanations

- "Even-If" -- Closer to decision boundary but do not cross it!
- Sub-type of CFs – counter to the facts, still an alternative world



Most Distant Neighbors

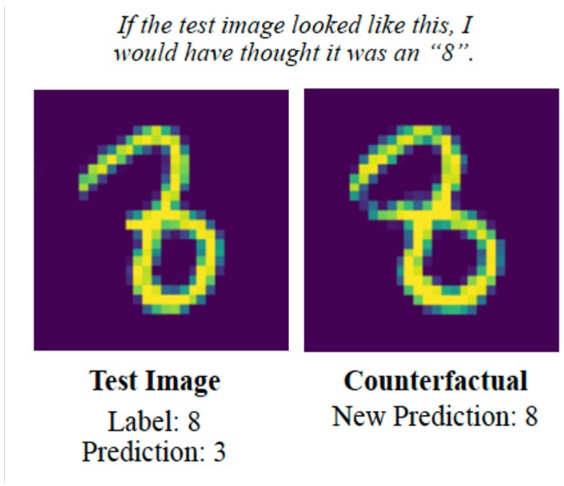
- *Baseline*: Find instance that is furthest from the query on any feature-dimension
- Analogous of NUNs (nb. actual data)
- Does not use counterfactual to guide search (aka CF-Free method)

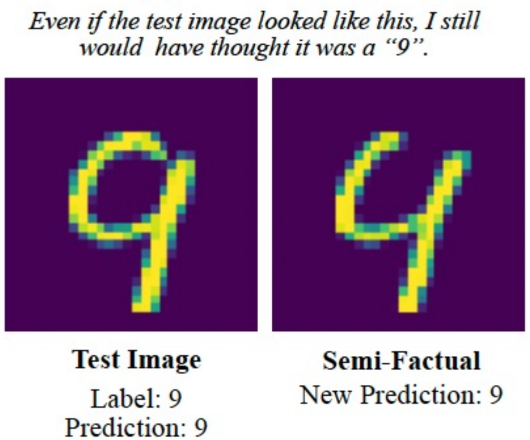


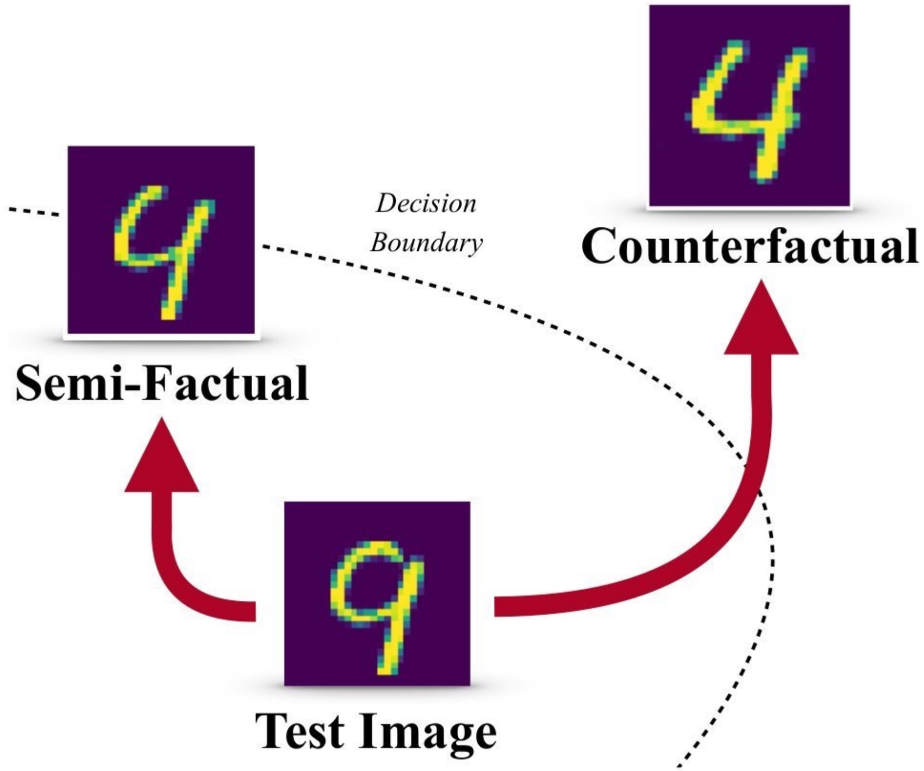
[Aryal & Keane, IJCAI-23]

PIECE

- PlausIble Exceptionality-based Contrastive Explanations
- Computes counterfactuals and semi-factuals

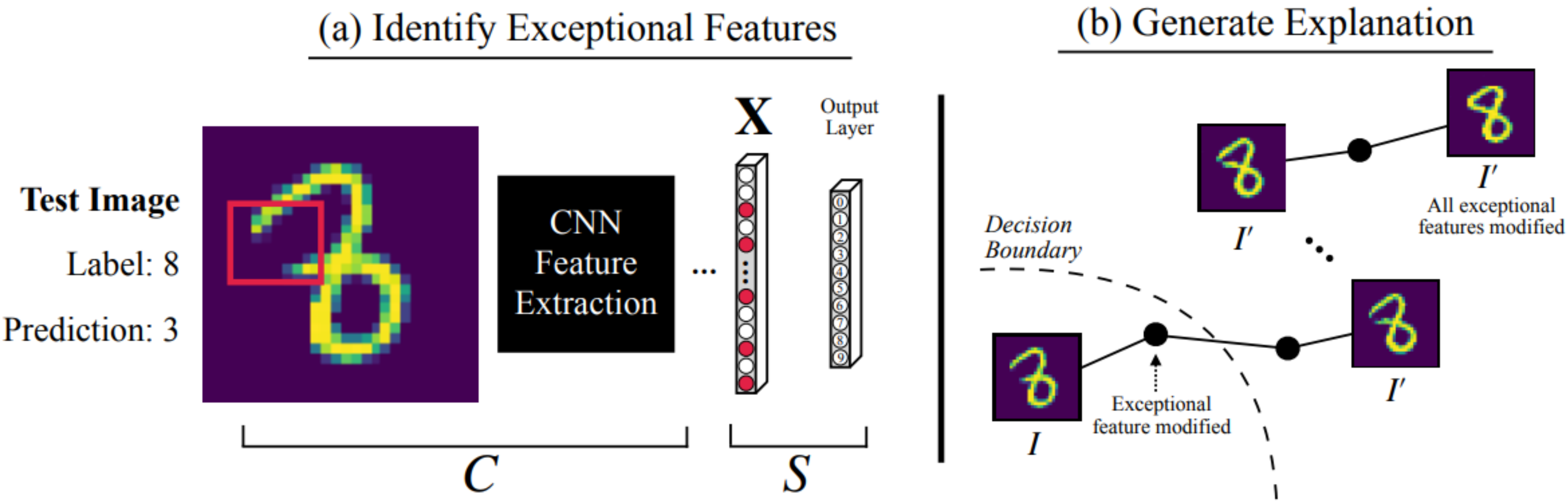






[Kenny & Keane AAAI-21]

PIECE



[Kenny & Keane AAAI-21]

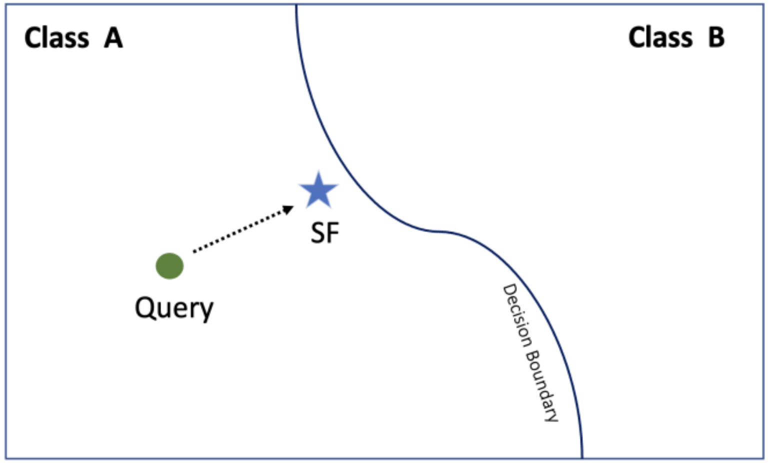
Computational Approaches

Counterfactual-Free

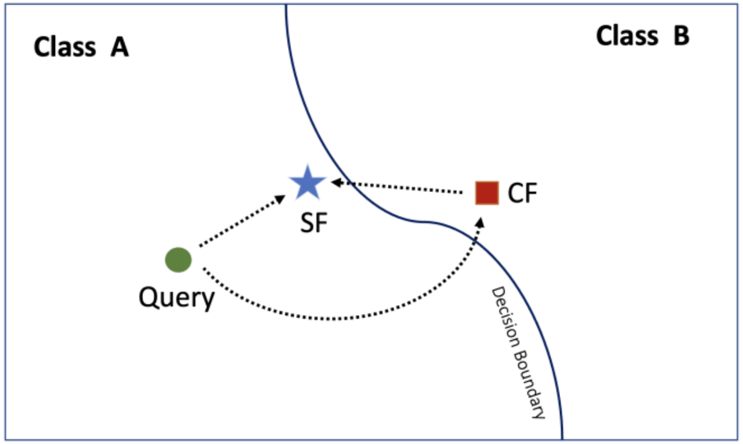
- Diverse Semifactual Explanations of Reject [Artelt 2022]
- Most Distant Neighbors [Aryal & Keane 23]

Counterfactual-Guided

- PIECE [Kenny & Keane AAAI-21]
- C2C-VAE [Zhao et al. ICCBR-22]



(a) Counterfactual-Free



(b) Counterfactual-Guided

Other & Current Topics

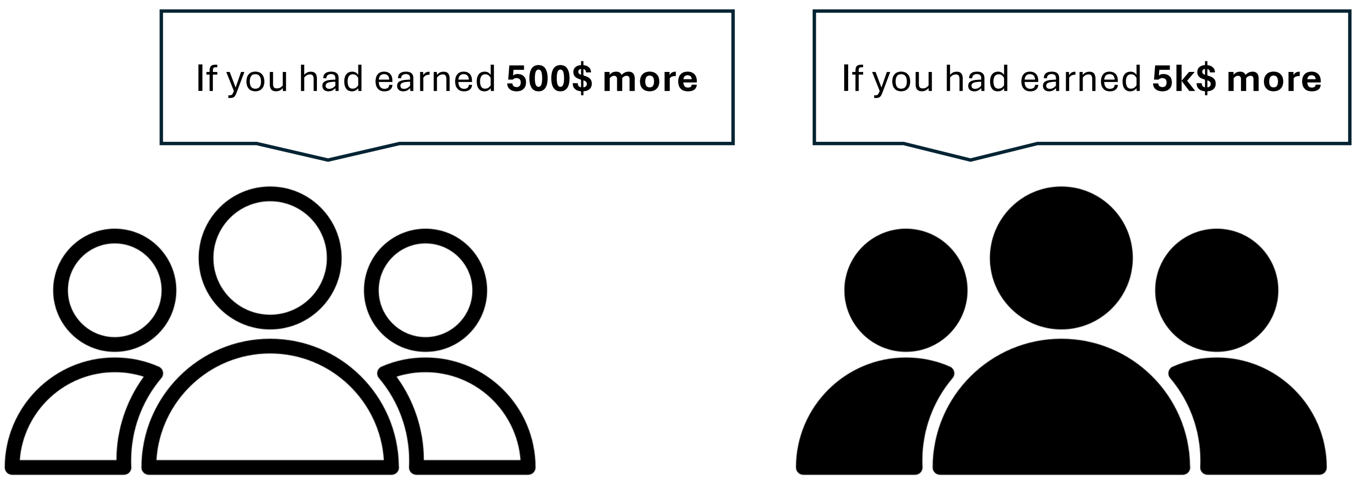
- Semi-Factuals
- Fairness
- Robustness
- Manipulations
- Group Counterfactuals

Fairness

- How can a CF be unfair?

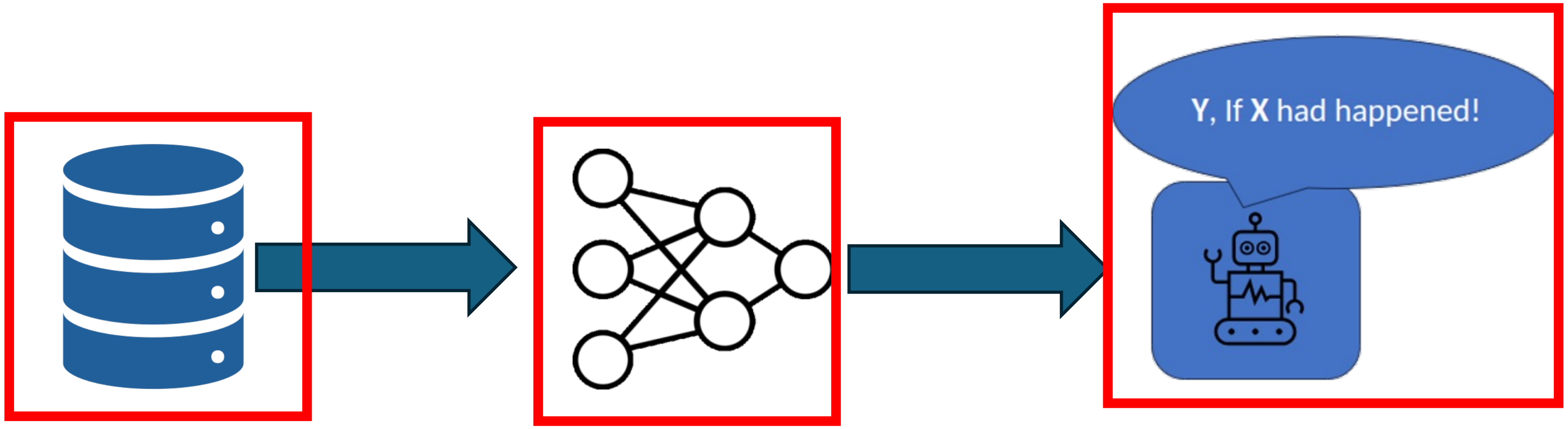
  => Differences in the cost of recourse (global vs. local)

- Group vs. Individual fairness [Slack 2021, Sharma 2020/21, Kügelgen 2022, Artelt 2023]

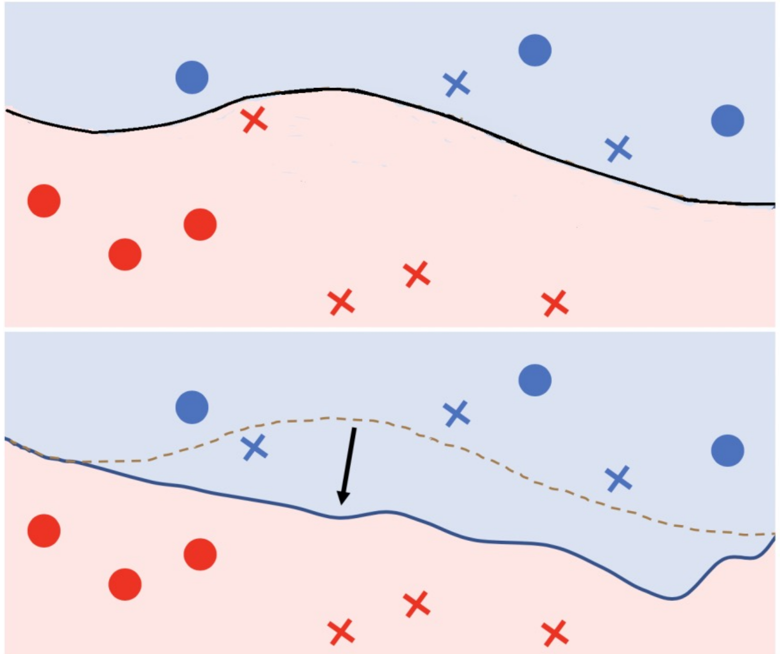


Approaches to Fix Fairness Issues

- Fix model [Sharma 2020/21]
- Fix CF generator [Kügelgen 2022, Artelt 2023]
- Fix data [Artelt 2024]



Make the Model Fair [Sharma 2020/21]



- Define cost of recourse for a group "g" as "Burden":

$$Burden(g) = \mathop{\mathbb{E}}_{g}[d(\mathbf{x}, \mathbf{c}^*)]$$

- Add fairness loss to training loss:

$$\mathcal{L}_{fairness} = \left| \mathop{\mathbb{E}}_{\mathbf{x}|\boxed{s(\boldsymbol{x})=a}}[d(\mathbf{x}, \mathcal{B})] - \mathop{\mathbb{E}}_{\mathbf{x}|\boxed{s(\boldsymbol{x})=b}}[d(\mathbf{x}, \mathcal{B})] \right|$$

Fix the CF generator [Artelt 2023]

- <u>Goal:</u> Same distributions of cost of recourse

$$p_{\vec{x}_{orig} \sim \mathcal{D}_1}\Big(\theta(\vec{x}_{orig}, CF(\vec{x}_{orig}, h))\Big) \approx p_{\vec{x}_{orig} \sim \mathcal{D}_2}\Big(\theta(\vec{x}_{orig}, CF(\vec{x}_{orig}, h))\Big)$$

- <u>Randomized algorithm:</u>

  => Sampe cost of recourse

$$\min_{\vec{x}_{\text{cf}} \in \mathbb{R}^d} \theta(\vec{x}_{\text{orig}}, \vec{x}_{\text{cf}}) + C_0 \cdot \ell\left(h(\vec{x}_{\text{cf}}), y_{\text{cf}}\right) + \boxed{C_1 \cdot \max\left(z - \theta(\vec{x}_{\text{orig}}, \vec{x}_{\text{cf}})\right)}$$

$$\text{where } z \sim p(\theta(\vec{X}_{\text{orig}}, \vec{X}_{\text{cf}}))$$

Fix the CF generator [Artelt 2023]

- Goal: Same

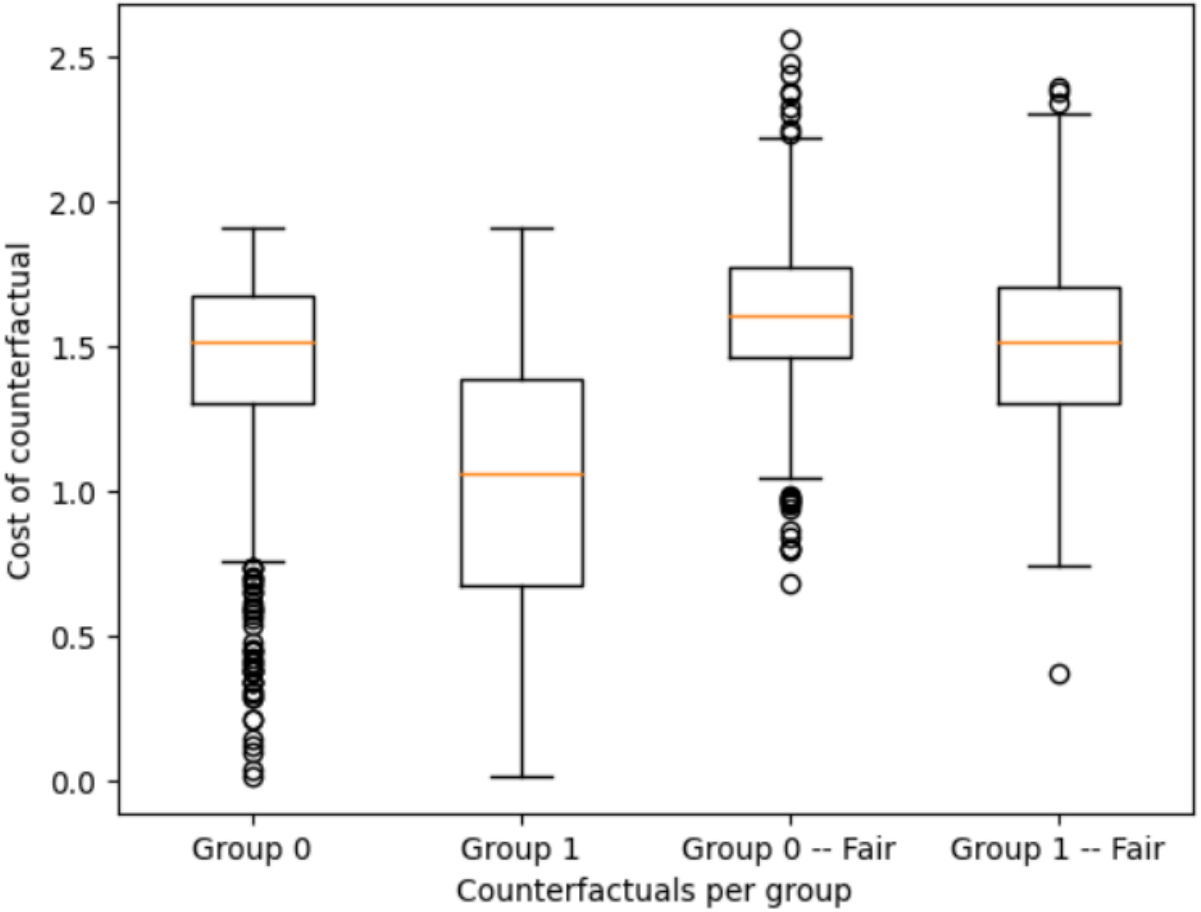
$$p_{\vec{x}_{orig}\sim\mathcal{D}_1}\Big(\theta(\vec{x}$$ $$))\Big)$$

- Randomized

=> Sampe cc

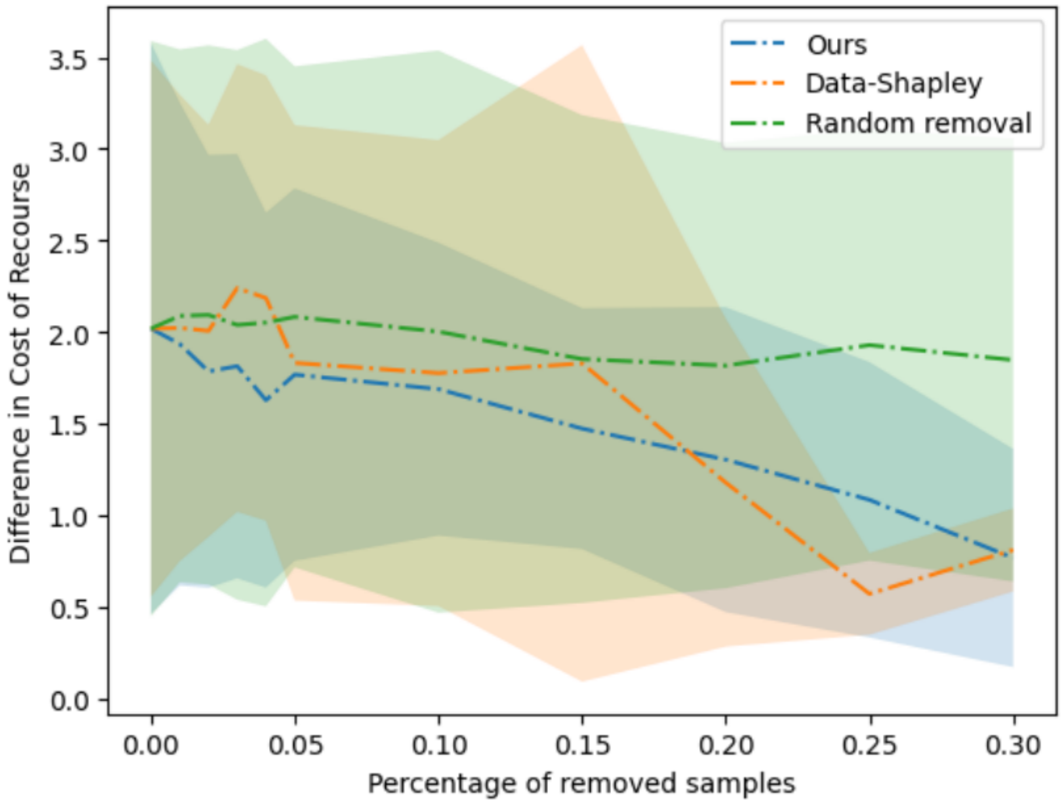
$$\min_{\vec{x}_{\mathrm{cf}}\in\mathbb{R}^d} \theta(\vec{x}_{\mathrm{orig}},$$ $$))$$

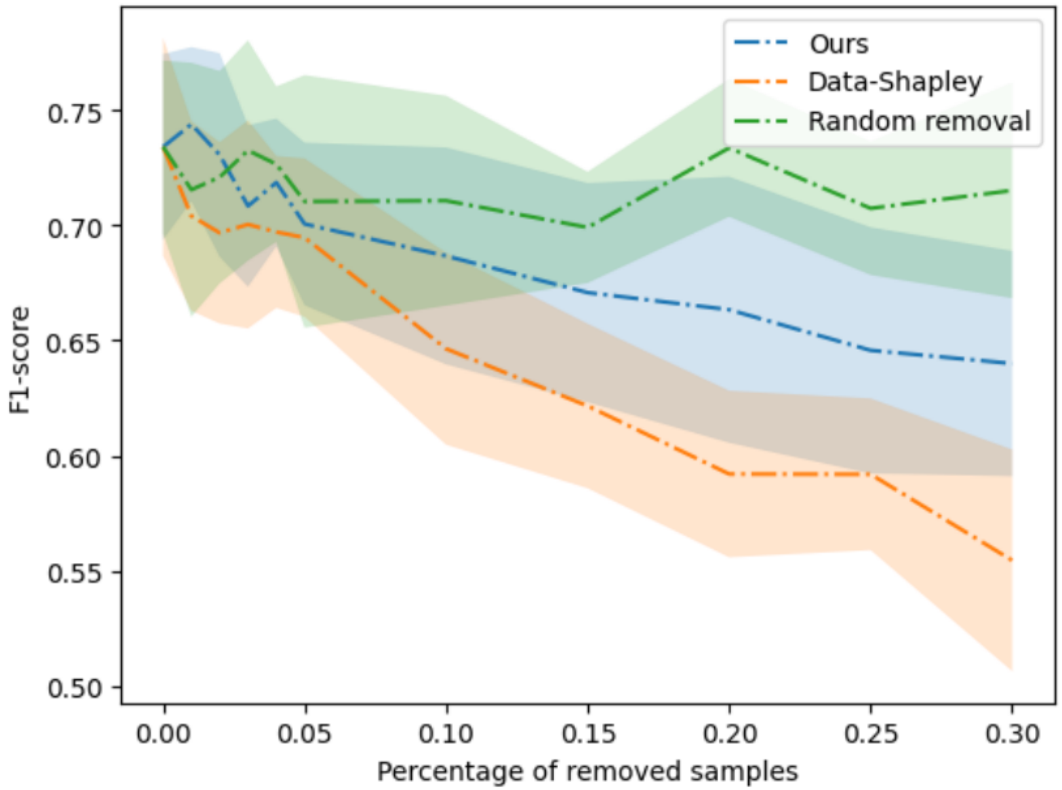
where $z \sim p(\theta($



Fix the Data [Artelt 2024]

- **Identify influential data points**





Open Questions 🤔

- Which approach to use?
  => Limitations?
- Formal guarantees?
  => Impossibility results?
- …

Other & Current Topics

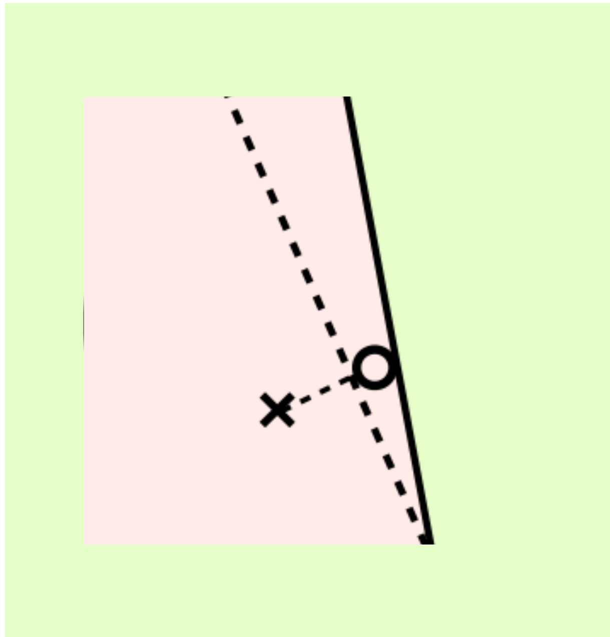
- Semi-Factuals
- Fairness
- Robustness
- Manipulations
- Group Counterfactuals

Robustness

- **(Missing) robustness** of CFs [Jiang 2024, Ferrario 2022, Artelt 2021]
    - Change in explanation
    - Change in cost
- *Input perturbations* (see individual fairness)
- *Model change* (over time)

=> Bad for the user!

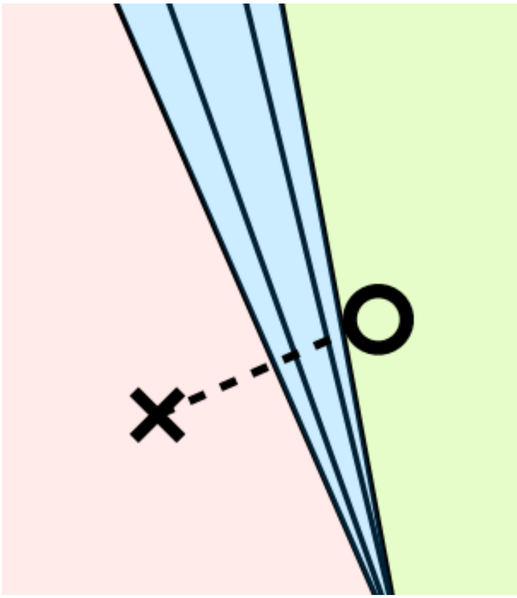
RQ: How to achieve robustness? 🤔



[Leofante 2023]

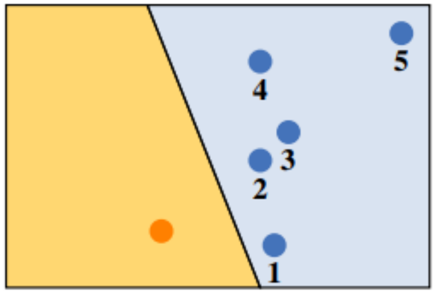
RQ: How to make CFs Robustness? 🤔

- Input changes:
  - Plausibility [Artelt 2020-21, Zhang 2023]
  - Min-Max Objective [Dominguez-Olmedo 2022]
  - Diversity [Leofante 2024]
- Model changes
  - Plausibility [Pawelczyk 2020]
  - Possible model changes -> MILPs [Jiang 2023-24]

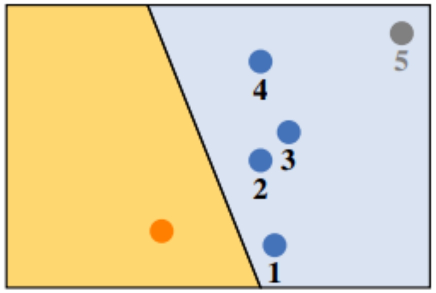


Robustness through Diversity [Leofante 2024]

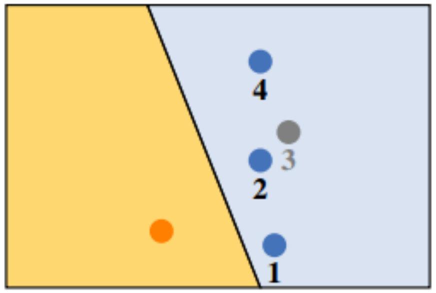
- Compute sets of **diverse CFs** instead of a single one
- **Nearby samples** should have large **overlap** in their diverse CFs!



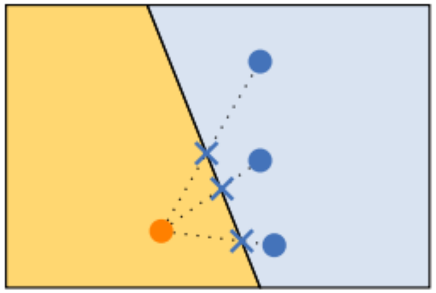
(a)



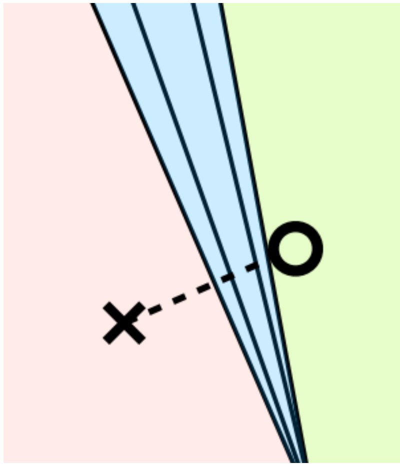
(b)



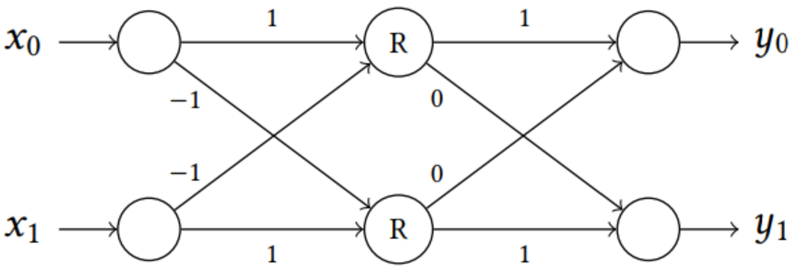
(c)

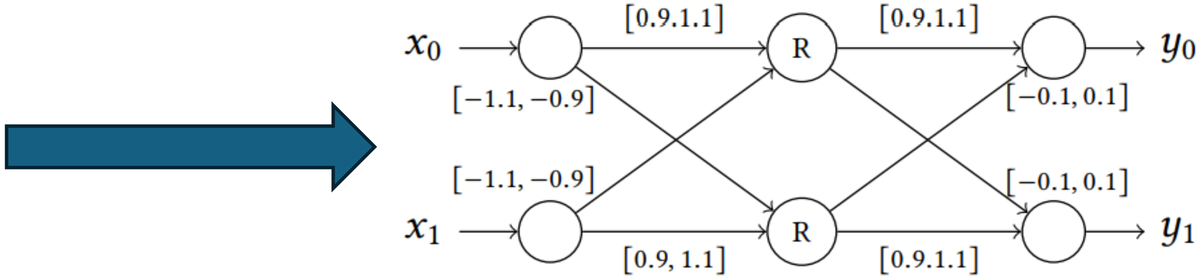

(d)

Robustness wrt. Model Changes [Jiang 2023-24]



- Consider ranges of parameters
- MILPs for computing & certifying robustness of CFs





Open Questions 🤔

- Cost of robustness
  - How much to sacrifice?
- Robustness and other aspects (e.g. plausibility, fairness, etc.)
  - Impossibility statements?
- ....

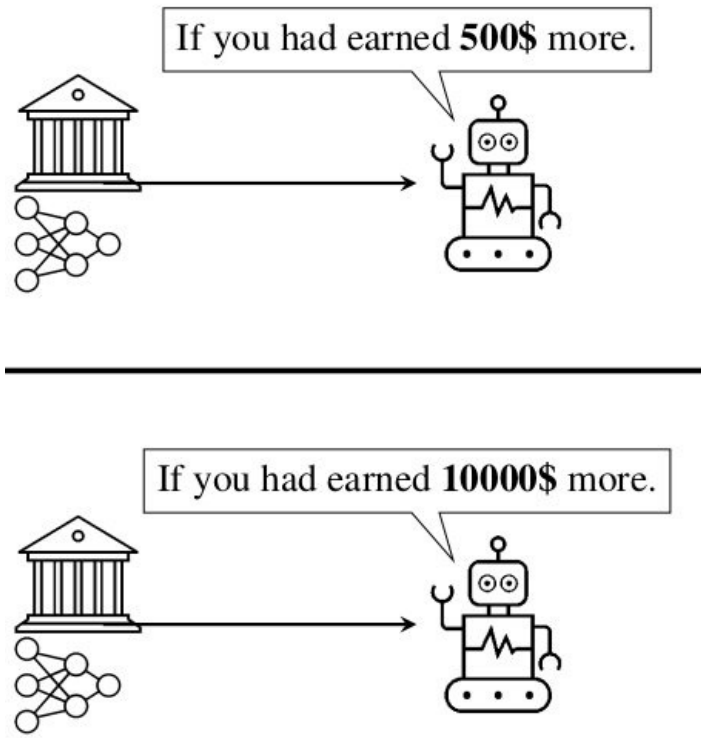
=> Checkout Survey on Robustness @ IJCAI [Jiang 2024]

Other & Current Topics

- Semi-Factuals
- Fairness
- Robustness
- Manipulations
- Group Counterfactuals

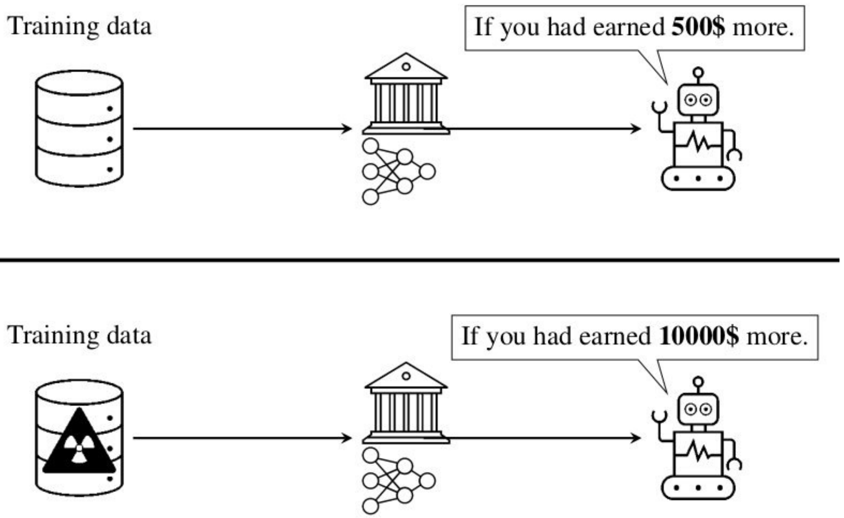
Manipulations

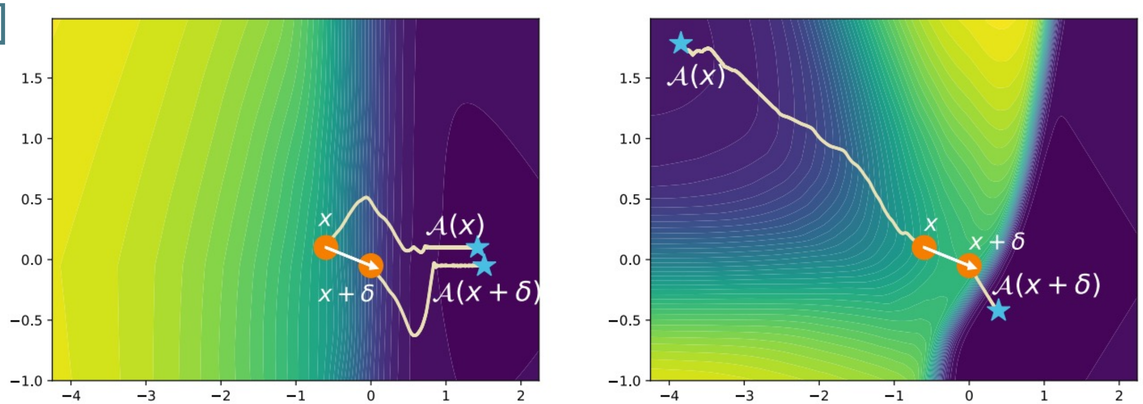
- Can CFs be manipulated?
  - Still trustworthy?
- Attack goals:
  - Cost of recourse
  - Fairness
- Manipulation on which level?



Manipulating CFs

- Adversarial training [Slack 2021]
  - Introduce "unfairness"
- Data poisoning [Artelt 2024]
  - Increase cost of recourse

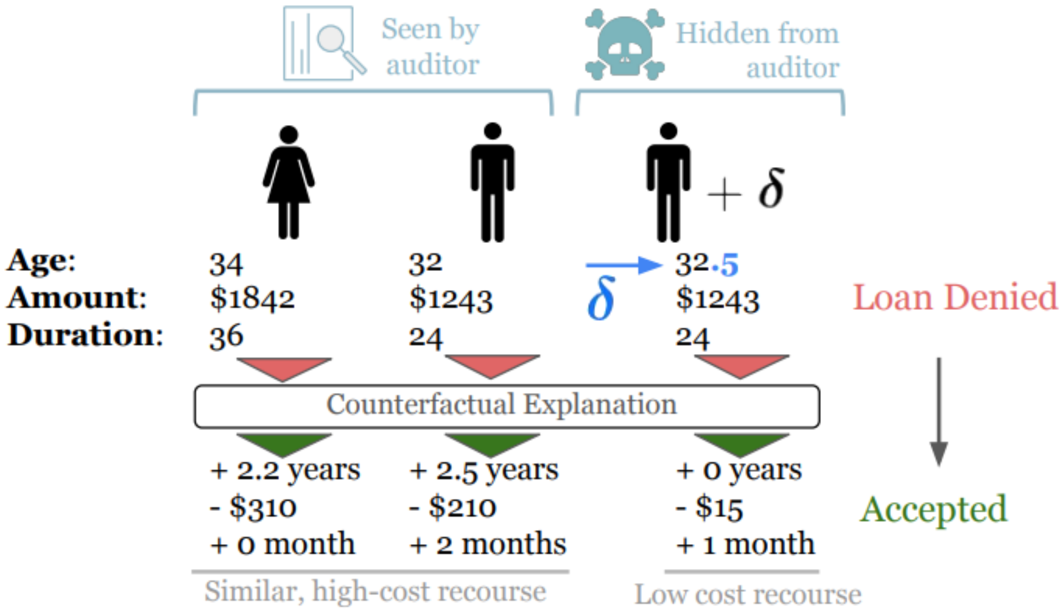




(a) **Training with BCE Objective** (b) **Training Adversarial Model**

Adversarial Training [Slack 2021]

- **"Backdoor":** Some small perturbation leads to lower cost recourse



Adversarial Training [Slack 2021]

- Assume: Attacker can manipulate training process

=> Adversarial training

Compute a CF

"Backdoor"

$$\mathbb{E}_{x\sim \boxed{\mathcal{D}_{\text{pr}}^{\text{neg}}}}\left[d\left(\boldsymbol{x}, \mathcal{A}(\boldsymbol{x})\right)\right] \gg \mathbb{E}_{x\sim \boxed{\mathcal{D}_{\text{np}}^{\text{neg}}}}\left[d\left(\boldsymbol{x}, \mathcal{A}(\boldsymbol{x}+\boldsymbol{\delta})\right)\right]$$

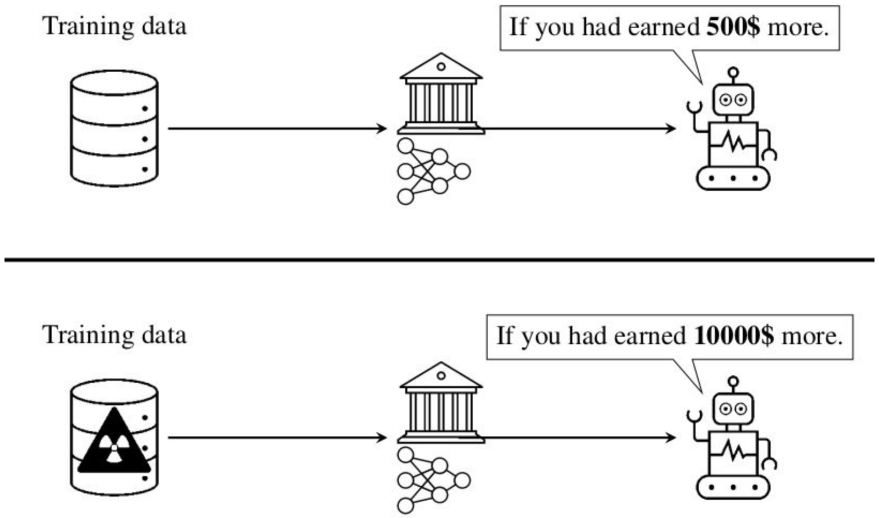
Table 2: **Recourse Costs of Manipulated Models**: Counterfactual algorithms find similar cost recourses for both subgroups, however, give much lower cost recourse if $\boldsymbol{\delta}$ is added before the search.

	Communities and Crime				German Credit			
	Wach.	**S-Wach.**	**Proto.**	**DiCE**	**Wach.**	**S-Wach.**	**Proto.**	**DiCE**
Protected	35.68	54.16	22.35	49.62	5.65	8.35	10.51	6.31
Non-Protected	35.31	52.05	22.65	42.63	5.08	8.59	13.98	6.81
Disparity	*0.37*	*2.12*	*0.30*	*6.99*	*0.75*	*0.24*	*0.06*	*0.5*
Non-Protected+$\boldsymbol{\delta}$	1.76	22.59	8.50	9.57	3.16	4.12	4.69	3.38
Cost reduction	*20.1×*	*2.3×*	*2.6×*	*4.5×*	*1.8×*	*2.0×*	*2.2×*	*2.0×*

Data Poisoning [Artelt 2024]

- Add poisonous samples to training data

=> Increase cost of recourse

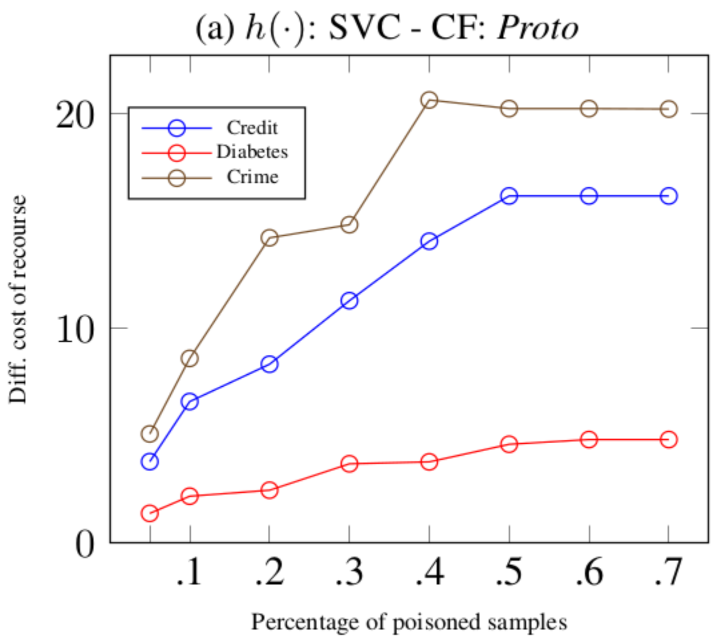


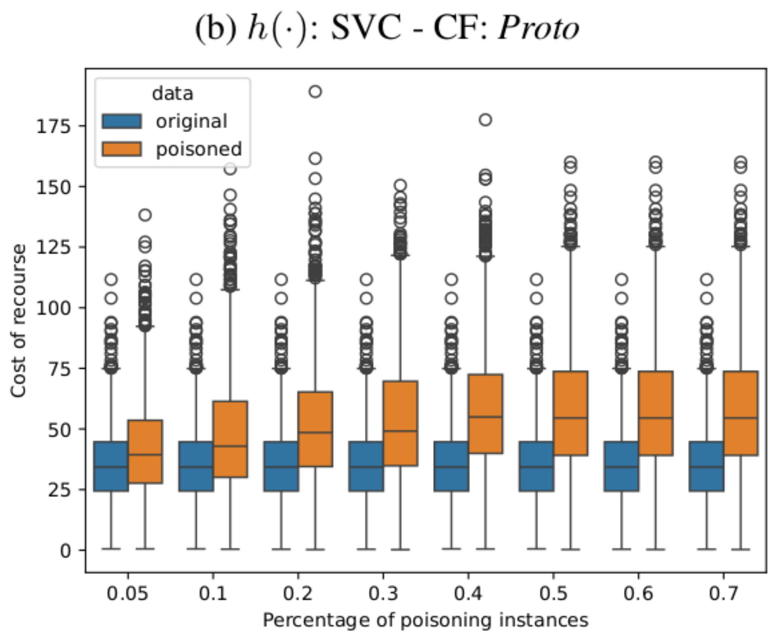
Data Poisoning [Artelt 2024]

- Finding poisonous samples as an optimization problem:

Poisc

Increase c
recourse





only

Defense against Manipulations

- Observation: Additional density constraints help [Artelt 2024]
- **Open questions!** 🤔

Other & Current Topics

- Semi-Factuals
- Fairness
- Robustness
- Manipulations
- Group Counterfactuals

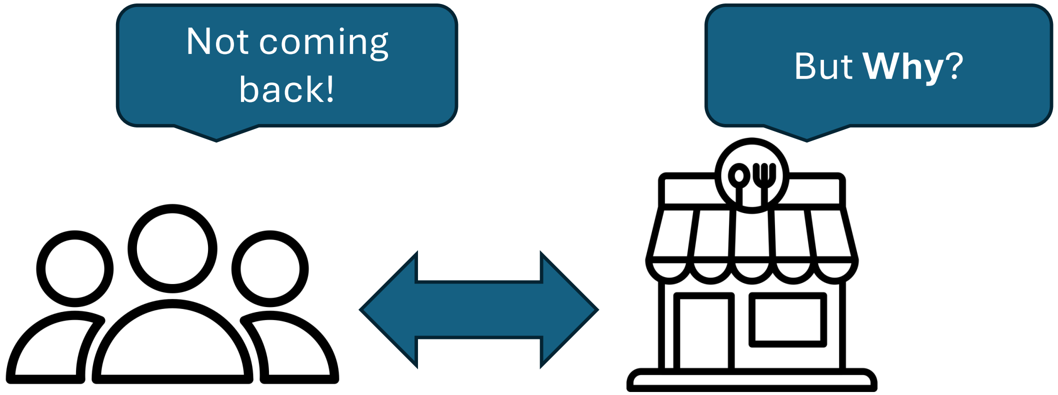
Group Counterfactuals

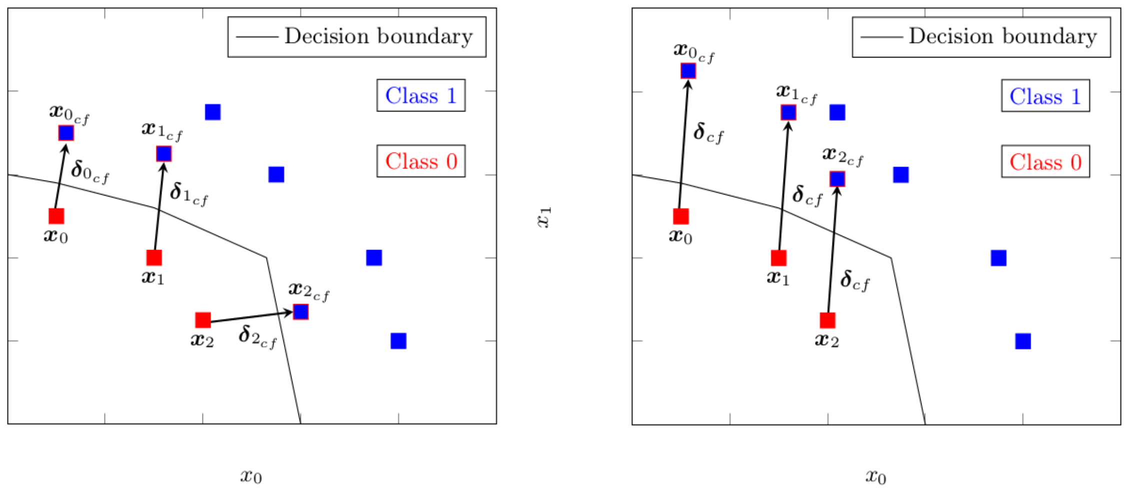
- One CF for many instances
- Real-world problems, e.g. Customer repurchase [Artelt 2024]

New challenges:

- Clustering/Grouping of instances

Current approaches [Kanamori 2022, Ley 2023, Warren 2023, Artelt 2023/24]





Group Counterfactuals

- Formalization [Artelt 2024]:

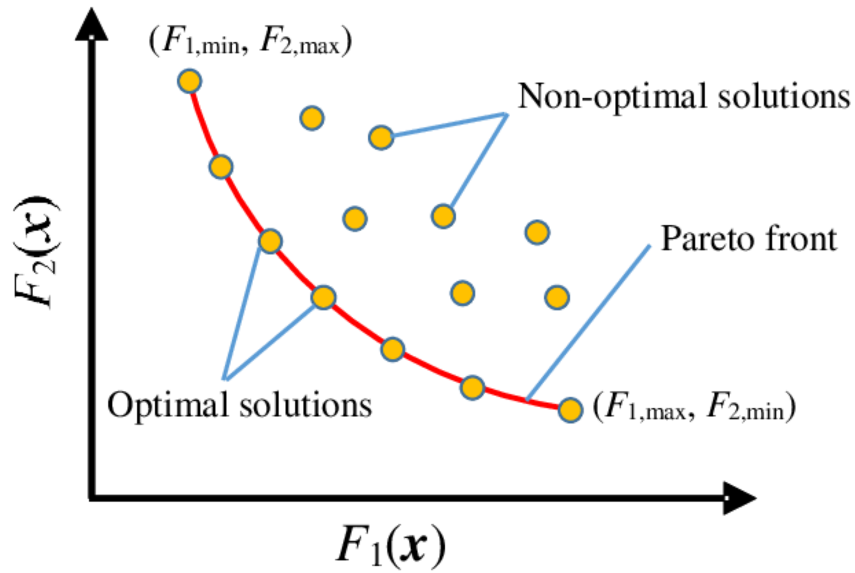
$$\min_{\vec{\delta}_{\text{cf}}\in\mathbb{R}^d} \left(\boxed{\theta(\vec{\delta}_{\text{cf}})}, \boxed{\sum_{\vec{x}_i\in\mathcal{D}} \ell(h(\vec{x}_i \oplus \vec{\delta}_{\text{cf}}), y_{\text{cf}})} \right)$$

Cost/Proximity **Contrastive**

- Feasibility?

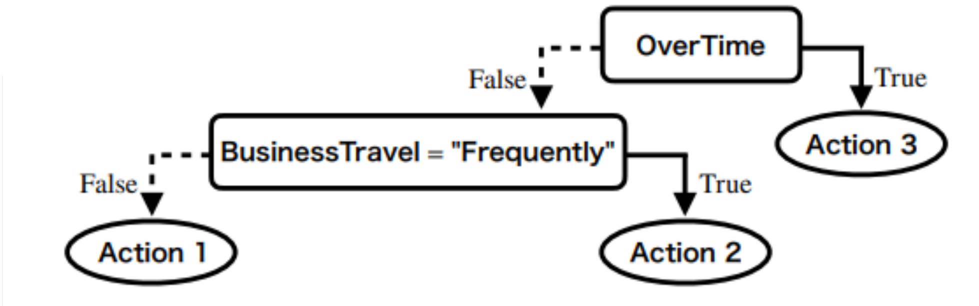
=> Pareto-optimal solutions!

What about other constraints?🤔



Group Counterfactuals

- Grouping given?
- Grouping and CFs
  - All-in-One [Kanamori 2022]
  - Separate [Artelt 2024]



	HowToChange	Effectiveness	
		Cost	Flip rate
Action 1	MonthlyIncome : + 1282$	0.17	83 %
Action 2	BusinessTravel : "Frequently" → "Rarely"	0.19	80 %
Action 3	OverTime : True → False	0.27	86 %

(a) Cluster with $\mathrm{d}(\vec{\delta}_{\mathrm{cf}i}, \vec{\delta}_{\mathrm{cf}j}) = \frac{\vec{\delta}_{\mathrm{cf}i}^{\top} \vec{\delta}_{\mathrm{cf}j}}{\|\vec{\delta}_{\mathrm{cf}i}\|_2 \|\vec{\delta}_{\mathrm{cf}j}\|_2}$

(b) Sub-cluster with $\mathrm{d}(\vec{\delta}_{\mathrm{cf}i}, \vec{\delta}_{\mathrm{cf}j}) = \|\theta(\vec{\delta}_{\mathrm{cf}i}) - \theta(\vec{\delta}_{\mathrm{cf}j})\|_2$

Open Questions 🤔

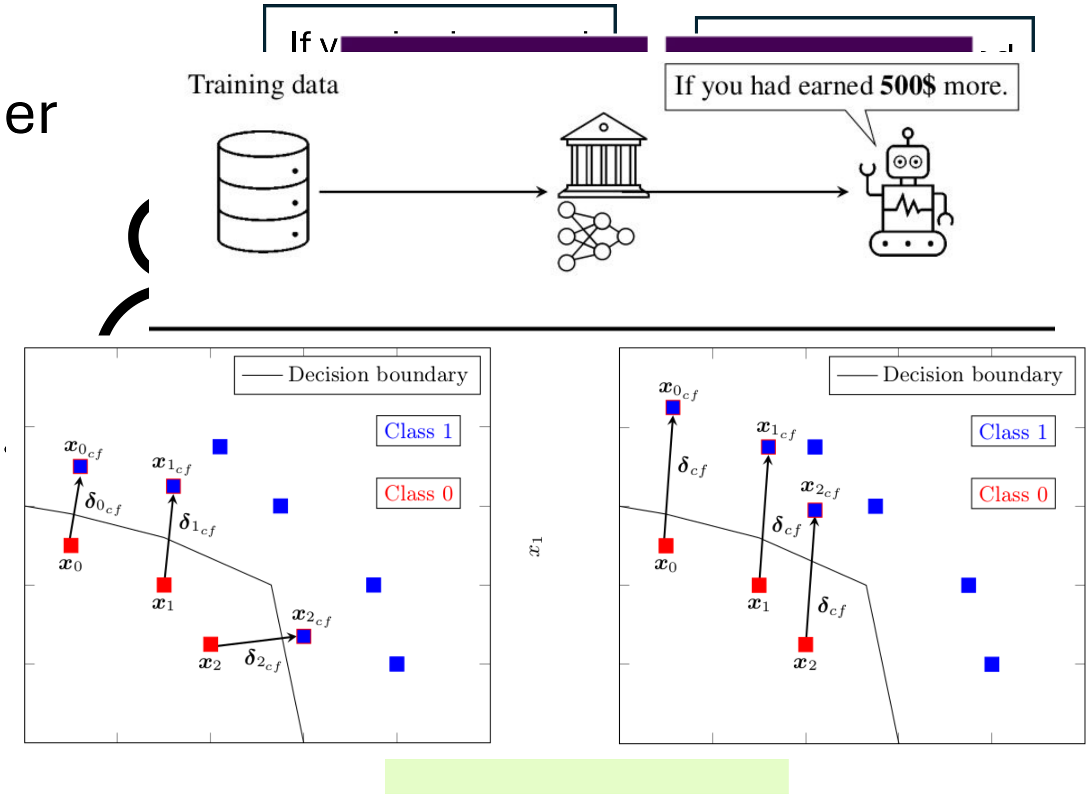
- Fairness?
- Robustness?

  => Outliers, poisonous instances, etc.
- Computational complexity and limitations?

  => Formal analysis
- Python framework/implementation
- ...

Summary

- **Robustness**
- **Fairness**
- **Semi-factuals** as the little brother of CFs
- CFs can be **manipulated**!
  - How to defend?
- CFs for groups of instances -- i.e. **group CFs**
  - Gaining popularity



Questions and Discussions

Ask us anything!